University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2012

Ranking of CTD articles and interactions using the OntoGene pipeline

Rinaldi, Fabio ; Clematide, Simon ; Hafner, Simon

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-62066
Conference or Workshop Item
Published Version

Originally published at:
Rinaldi, Fabio; Clematide, Simon; Hafner, Simon (2012). Ranking of CTD articles and interactions using the OntoGene pipeline. In: 2012 BioCreative workshop, Washington D.C., 4 April 2012 - 5 April 2012.

Ranking of CTD articles and interactions using the OntoGene pipeline

Fabio Rinaldi, Simon Clematide and Simon Hafner
Institute of Computational Linguistics, University of Zurich
{rinaldi,siclemat}@cl.uzh.ch,hafnersimon@gmail.com

Abstract

In this paper we briefly describe the architecture of the OntoGene Relation mining pipeline and its application in the task 1 of BioCreative IV. The aim of the task is to deliver information useful for the triage of abstracts relevant to the process of curation of the Comparative Toxicogenomics Database.

Although the main focus of our text mining research is the extraction of interactions, we decided to participate in the task with the assumption that articles which contain relevant interactions would be relevant themselves.

We use a conventional information retrieval system (Lucene) to provide a baseline ranking, which we then combine with information provided by the relation mining module, in order to achieve an optimized ranking.

1 Introduction

As a way to cope with the constantly increasing generation of results in molecular biology, some organizations maintain various types of databases that aim at collecting the most significant information in a specific area. For example, UniProt/SwissProt [14] collects information on all known proteins. IntAct [4] is a database collecting protein interactions. The Comparative Toxicogenomics Database collects interactions between chemicals and genes in order to support the study on the effects of environmental chemicals on health [6]. Most of the information in these databases is derived from the primary literature by a process of manual revision known as "literature curation". Text mining solutions are increasingly requested to support the process of curation of biomedical databases.

The work presented here is part the OntoGene project[1], which aims at improving biomedical text mining through the usage of advanced natural language processing techniques. Our approach relies upon information delivered by a pipeline of NLP tools, including sentence splitting, tokenization, part of speech tagging, term recognition, noun and verb phrase chunking, and a dependency-based syntactic analysis of input sentences [11, 9]. The results of the entity detection feed directly into the process of identification of interactions. The syntactic parser [13] takes into account constituent boundaries defined by previously identified multi-word entities. Therefore the richness of the entity annotation has a direct beneficial impact on the performance of the parser, and thus leads to better recognition of interactions.

Recently, in the context of the SASEBio project (Semi-Automated Semantic Enrichment of the Biomedical Literature), the OntoGene group has developed a user-friendly interface (ODIN: OntoGene Document INspector) which presents the results of the text mining pipeline intuitive fashion, and allows a deeper interaction of the curator with the underlying text mining system.

In the rest of this paper we first explain how our existing OntoGene relation mining system has been customized for CTD (section 2) and then how it has been integrated with a conventional IR system (Lucene) for the purpose of the Triage task (section 3). We also provide a brief overview of our ODIN curation interface (section 4) and a preliminary evaluation of the results obtained so far (section 5)

[1] http://www.ontogene.org/

2 Methods

In this section we describe the OntoGene Text Mining pipeline which is used to (a) provide all basic preprocessing (e.g. tokenization) of the target documents, (b) identify all mentions of domain entities and normalize them to database identifiers, and (c) extract candidate interactions. We describe then in some detail a machine learning approach used to obtain an optimized scoring of candidate interactions based upon global information from the whole CTD.

2.1 Preprocessing and Detection of Domain Entities

In order to solve the triage task, we processed the PubMed abstracts of the referenced articles by the OntoGene Text Mining pipeline.
As shown in our previous work [2], the inclusion of PubMed metadata, such as the list of chemical substances, as well as the annotated MeSH descriptors and qualifiers, improves the detection of important relations and enhances term recognition coverage. Therefore, we added these metadata from the PubMed XML files as a textual list at the end of each abstract. In our OntoGene text mining pipeline, the sentence and token boundaries of the enriched abstracts are identified using the LingPipe framework[2]
Next, we describe our approach to the problem of detecting names of relevant domain entities in biomedical literature (genes, chemicals and diseases for CTD) and grounding them to widely accepted identifiers assigned by the original database. Terms, i.e. preferred names and synonyms, are automatically extracted from the original CTD databases and stored in a common internal format, together with their unique identifiers (as obtained from the original resource). An efficient lookup procedure is used to annotate any mention of a term in the documents with the ID(s) to which it corresponds. A term normalization step is used to take into account a number of possible surface variations of the terms. The same normalization is applied to the list of known terms at the beginning of the annotation process, when it is read into memory, and to the candidate terms in the input text, so that a matching between variants of the same term becomes possible despite the differences in the surface strings. In case the normalized strings match exactly, the input sequence is annotated with the IDs of the reference term – no further disambiguation on concepts is done. For more technical details of our term recognizer, see [8].

2.2 Detection of Interactions

The information about mentions of relevant domain entities (and their corresponding unique identifiers) can be used to create candidate interactions. In other words, the co-occurrence of two entities in a given text span (typically one or more sentences, or an even larger observation window) is a low-precision, but high-recall indication of a potential relationship among those entities. In order to obtain better precision it is possible to take into account the syntactic structure of the sentence, or the global distribution of interactions in the original database. In this section we describe in detail how candidate interactions are ranked by our system, according to their relevance for CTD curation, by exploiting the vast amount of curated articles in the CTD base.
An initial ranking of the candidate relations can be generated on the basis of frequency of occurrence of the respective entities only:

$$relscore(e_1, e_2) = (f(e_1) + f(e_2))/f(E)$$

where $f(e_1)$ and $f(e_2)$ are the number of times the entities e_1 and e_2 are observed in the abstract, while $f(E)$ is the total count of all identifiers in the abstract. We know from our previous experiments [9] that giving a "boost" of 10 to the entities contained in the title produces a measurable improvement of ranking of the results. This simple approach can be further optimized if we apply a supervised machine learning method for scoring the probability of a term to be part of an interesting relation. There are two key motivations for scoring concepts based upon relation candidate ranking: First, we need to adapt to highly-ranked false positive relations which are generated by a simple frequency based approach by frequent but uninteresting concepts. The goal is to model some global properties of the curated CTD relations. Second, we want to penalize false positive concepts that our term recognizer detects. In order to deal with such cases, we need to condition the concepts by their normalized[3] textual form t. The combination of a term t and one of its valid entities e is noted as $t:e$.

[2]More information regarding the framework can be found at `http://alias-i.com/lingpipe`.
[3]A normalized textual form of a term consists of the sequence of lower-case alphanumeric characters of all term tokens.

Next we define a predicate $gold(A, e)$ which is true (i.e. 1) for an article A if there is at least one relation in the gold standard where entity e is part of, and false (i.e. 0) otherwise. We estimate the overall probability $P(\, gold(A, e) = 1 \mid t\!:\!e\,)$ with the help of the Maximum Entropy Modeling tool *megam* [3]. For training we use the set of CTD-referenced PubMed articles having not more than 12 manually curated relations[4], additionally removing all articles which are part of the BioCreative training and test set for the respective data set[5].
For unseen normalized terms t, i.e. terms not present in the training data, the maximum entropy classifier would assign a low default probability based on the distribution of all training instances. However, we can specify better back-off probabilities if we take into account the admissible entity/entities e of term t. Our current back-off model works as follows: if the entity e of an unseen t is seen in the article, the averaged probability of all seen term-entity pairs is used. Otherwise, the averaged probability of all entities of the same type as e is used.
The score of an entity e in an article A is the sum of all zoned term frequencies[6] weighted by their gold probability:

$$score(e) = \sum_{t:e \in A} f(t\!:\!e) \times P(\, gold(A, e) = 1 \mid t\!:\!e\,)$$

Having determined the individual score for each entity e, we compute the relation score as the harmonic mean of its component scores:

$$relscore(e_1, e_2) = 2 \times \frac{score(e_1) \times score(e_2)}{score(e_1) + score(e_2)}$$

In preceding work on relation ranking [2], the relation score was taken as a sum of the concept scores. By performing systematic cross-validation experiments on all CTD articles, we noticed that using the harmonic mean improves the results considerably. In order to make the relation scores comparable between different articles we normalize all relation for a given BioCreative data set.

3 Integration with a standard IR system

A conventional IR system is used to provide a baseline document classification. Information derived from the OntoGene pipeline, and from the ranking process described in the previous section, is then added as additional features in order to improve the baseline ranking generated by the IR system. The integration of the various components is performed using mainly JRuby (and some small parts in Java).

3.1 Terminology-aware tokenization

Documents are processed by Lucene in the conventional way, selecting different boost values for title and abstract (10 for title, 3 for abstract, just as in the CTD reference system). The Lucene API is accessed via JRuby. Changes in the boost values did not show any statistically significant change in the MAP scores, because most of the information is in the abstract, not the title. The existence of relevant information in the title typically implies relevant information in the abstract.
The only significant technical change to Lucene preprocessing is the replacement of the "StandardAnalyzer" component (which is the default analyzer for English, responsible for tokenization, stemming, etc.) with our own tokenization results, as delivered by the OntoGene pipeline. The advantage of this approach is that we can flexibly treat recognized technical terms as individual tokens, and map together their synonyms [7]. In order words, after this step all known synonyms of a term will be treated as identical by the IR system.
The "StandardAnalyzer" component is replaced by a simple transformation of the XML output of the pipeline into a format suitable for internal processing by Lucene. In particular tokens and terms as recognized by the pipeline are transformed into Lucene "token" data objects. Whenever a domain entity (denoted by the `Term` element in the XML representation) is found, its words are concatenated to one token. At the same position, a new token with the text of the concept identifier is added to the stream.
As an example:

[4] The threshold of 12 relations is motivated by the observation that the more relations an article has the less probable it is to find them by processing the abstracts only.
[5] This results in 22319 articles for the BioCreative 4 training set, containing 69320 curated relations. For the BioCreative 4 test set, we used 22825 articles with 71064 relations.
[6] As mentioned earlier, occurrences in the title are counted 10 times.

```
<W C="VBN" id="W151" o1="758" o2="767">inhibited</W>
<Term allvalues="MESH_D015232:chem" id="TW152W153"
      matched="prostaglandine2" type="chem">
  <W C="NN" id="W152" o1="768" o2="781">prostaglandin</W>
  <W C="NN" id="W153" o1="782" o2="784">E2</W>
</Term>
<W C="NN" id="W154" o1="785" o2="794">synthesis</W>
```

would be converted to the following (square brackets denote token boundaries):

```
[inhibited] [prostaglandin E2] [synthesis]
            [MESH_D015232]
```

Synonymous terms (as identified by the pipeline) are mapped to their unique identifiers (for this experiment the term identifier provided by the CTD database, which happens to be a MeSH term in the example above). A basic search is conducted by mapping the target chemical to the corresponding identifier, which is then used as a query term to perform a search in Lucene.

3.2 Relation-based query expansion

As described in section 2.2 the OntoGene pipeline is not only used in order to deliver an optimized tokenization, it can also be used to generate candidate interactions, which could be directly used for curation purposes by CTD curators.
Although the definition of the task did not require the participants to deliver candidate interactions, we worked under the assumption that documents which contain relevant interactions would be relevant themselves. When another term is often seen in relation with the target term, it is probably important for the target. This statistical information is used to adjust the ranking of the documents.
The OntoGene pipeline delivers candidate interactions as part of its standard output for each single document. Each interaction is assigned a score in the interval (0,1]. The relations are extracted from all the files in the document set assigned to the target chemical by the organizers. All relations which involve a term equivalent to the target (the target or one of its synonyms) are extracted. The interacting entity (the second term in those interactions) is then added to the search query, for each interaction, giving rise to an expanded query. The additional query terms are weighted according to the normalized score of the original interaction. As an example suppose two documents contain the interactions listed in the first two columns below (document 1 and document 2):

```
document 1:    document 2:       expansion terms:
A C 1          A B 1             C 1    from doc 1
B C 0.7        B D 0.42          B 0.75 from doc 1 (score 0.5) and doc 2 (score 1)
A B 0.5                          D 0.4  from doc 1
A D 0.4
```

If the target term is A, the relations marked in boldface are relevant, which gives us new search terms to be added to the query, listed in the 3rd column with their normalized weights (sum of scores divided by the number of relations). The original target term is given a weight which is above the weight of the relations in order to make it clearly more relevant than any of the added terms. We have experimentally verified on the training data that the using this query expansion process improves the average MAP scores from 0.62225 to 0.694625 (i.e. an improvement of nearly 12%).

4 The ODIN Interface

The results of the OntoGene text mining system are made accessible through a curation system called **ODIN** ("OntoGene Document INspector") which allows a user to dynamically inspect the results of their text mining pipeline. A previous version of ODIN was used for participation in the 'interactive curation' task (IAT) of the BioCreative III competition [1]. This was an informal task without a quantitative evaluation of the participating systems. However, the curators who used the system commented extremely positively on its usability for a practical curation task. An experiment in interactive curation has been performed in collaboration with curators of the PharmGKB database [5, 12]. The results of this experiment are described in [10], which also provides further details on the architecture of the system.

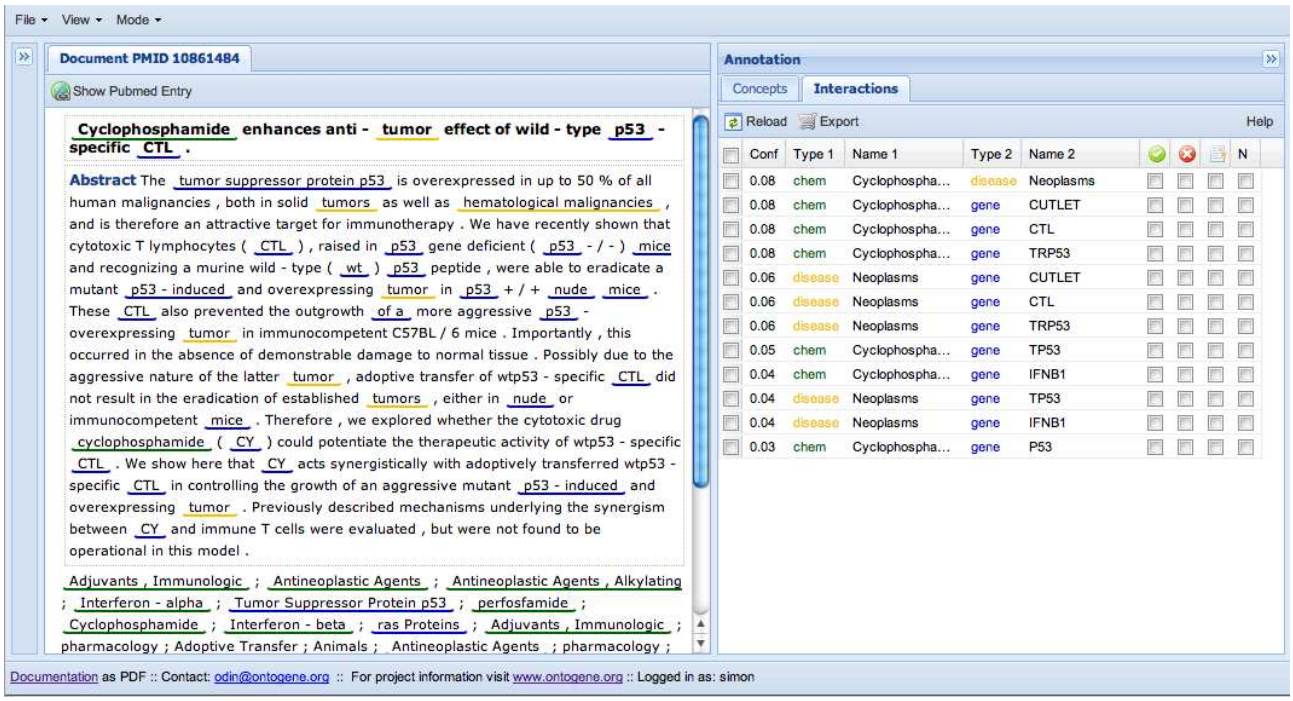

Figure 1: Entity annotations and candidate interactions on a sample PubMed abstract

More recently, we partially adapted ODIN to the aims of CTD curation, allowing the inspection of PubMed abstracts annotated with CTD entities and showing the interactions extracted by our system. We would be interested in providing further customizations according to the needs of the CTD curation process.

5 Evaluation

In order to generally assess the upper limit of our relation recognition system, we evaluated the coverage of the term recognizer on all CTD-referenced articles containing at most 12 curated relations. The table below describes the coverage of term recognition for concepts and relations in experimental data and shows that we find about 3/4 of all entities. However, the upper limits for relation detection are not the same for all relation types. The coverage of relations involving chemicals have substantially lower coverage rates which seems a bit unfortunate for the CTD triage task.

Cat	N	abs	rel
disease	12639	9502	75.18
chemical	38523	30129	78.21
gene	39150	29199	74.58
TOTAL	90312	68830	76.21
dis-gen	6956	5126	73.69
che-dis	12154	8356	68.75
che-gen	52746	34883	66.13
TOTAL	71856	48365	67.31

The table below shows the final results obtained on the training set using the on-line evaluation tool. Due to lack of space and time, we cannot report here a detailed analysis of all intermediate results, which we intend to present at the workshop.

Term	MAP	genes	chemicals	diseases
doxorubicin	0.800	0.167	0.843	0.793
indomethacin	0.936	0.331	0.834	0.725
raloxifene	0.798	0.244	0.818	0.778
amsacrine	0.655	0.603	0.689	0.500
aniline	0.543	0.625	0.561	0.524
2-Acetylaminofluorene	0.643	0.412	0.845	0.421
aspartame	0.365	0.686	0.756	0.720
quercetin	0.853	0.463	0.646	0.653

6 Conclusions

In this paper we have described our approach towards ranking biomedical abstracts for the triage task of the CTD curation process. The peculiarity of the approach is that it gives priority to the identification of candidate interactions, which are then used as additional weighting factors in a conventional IR-based system.
The OntoGene pipeline is capable of delivering all information relevant to CTD curation: entities with their database references, interactions, and interaction terms. In the shared task however, due to insufficient time for customization, we decided to exclude the computation of interaction terms. The results of the system are accessible through an intuitive interactive interface, which we are willing to customize for CTD curation.

Acknowledgments
This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1) and Novartis Pharma AG, NIBR-IT, Text Mining Services, CH-4002, Basel, Switzerland.

References

[1] Cecilia Arighi, Phoebe Roberts, Shashank Agarwal, Sanmitra Bhattacharya, Gianni Cesareni, Andrew Chatr-aryamontri, Simon Clematide, Pascale Gaudet, Michelle Giglio, Ian Harrow, Eva Huala, Martin Krallinger, Ulf Leser, Donghui Li, Feifan Liu, Zhiyong Lu, Lois Maltais, Naoaki Okazaki, Livia Perfetto, Fabio Rinaldi, Rune Saetre, David Salgado, Padmini Srinivasan, Philippe Thomas, Luca Toldo, Lynette Hirschman, and Cathy Wu. Biocreative iii interactive task: an overview. *BMC Bioinformatics*, 12(Suppl 8):S4, 2011.

[2] Simon Clematide and Fabio Rinaldi. Ranking interactions for a curation task. *Machine Learning and Applications, Fourth International Conference on*, 2:100–105, 2011.

[3] Hal Daumé III. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`, August 2004.

[4] Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, Hanah Margalit, John Armstrong, Amos Bairoch, Gianni Cesareni, David Sherman, and Rolf Apweiler. IntAct: an open source molecular interaction database. *Nucl. Acids Res.*, 32(suppl 1):D452–455, 2004.

[5] T.E. Klein, J.T. Chang, M.K. Cho, K.L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D.E. Oliver, D.L. Rubin, F. Shafa, J.M. Stuart, and R.B. Altman. Integrating genotype and phenotype information: An overview of the pharmgkb project. *The Pharmacogenomics Journal*, 1:167–170, 2001.

[6] C.J. Mattingly, M.C. Rosenstein, G.T. Colby, J.N. Forrest Jr, and J.L. Boyer. The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *Journal of Experimental Zoology Part A: Comparative Experimental Biology*, 305A(9):689–692, 2006.

[7] Fabio Rinaldi, James Dowdall, Michael Hess, Kaarel Kaljurand, Mare Koit, Kadri Vider, and Neeme Kahusk. Terminology as Knowledge in Answer Extraction. In *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE02)*, pages 107–113, Nancy, 28–30 August 2002.

[8] Fabio Rinaldi, Kaarel Kaljurand, and Rune Saetre. Terminological resources for text mining over biomedical scientific literature. *Journal of Artificial Intelligence in Medicine*, 52(2):107–114, June 2011.

[9] Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, and Therese Vachon. OntoGene in BioCreative II. *Genome Biology*, 9(Suppl 2):S13, 2008.

[10] Fabio Rinaldi, Gerold Schneider, and Simon Clematide. Relation mining experiments in the pharmacogenomics domain. *Journal of Biomedical Informatics, special issue for the Biocuration 2012 conference*, 2012. conditionally accepted for publication.

[11] Fabio Rinaldi, Gerold Schneider, Kaarel Kaljurand, Michael Hess, and Martin Romacker. An Environment for Relation Mining over Richly Annotated Corpora: the case of GENIA. *BMC Bioinformatics*, 7(Suppl 3):S3, 2006.

[12] Katrin Sangkuhl, Dorit S. Berlin, Russ B. Altman, and Teri E. Klein. Pharmgkb: Understanding the effects of individual genetic variants. *Drug Metabolism Reviews*, 40(4):539–551, 2008. PMID: 18949600.

[13] Gerold Schneider. *Hybrid Long-Distance Functional Dependency Parsing*. Doctoral Thesis, Institute of Computational Linguistics, University of Zurich, 2008.

[14] UniProt Consortium. The universal protein resource (uniprot). *Nucleic Acids Research*, 35:D193–7, 2007.